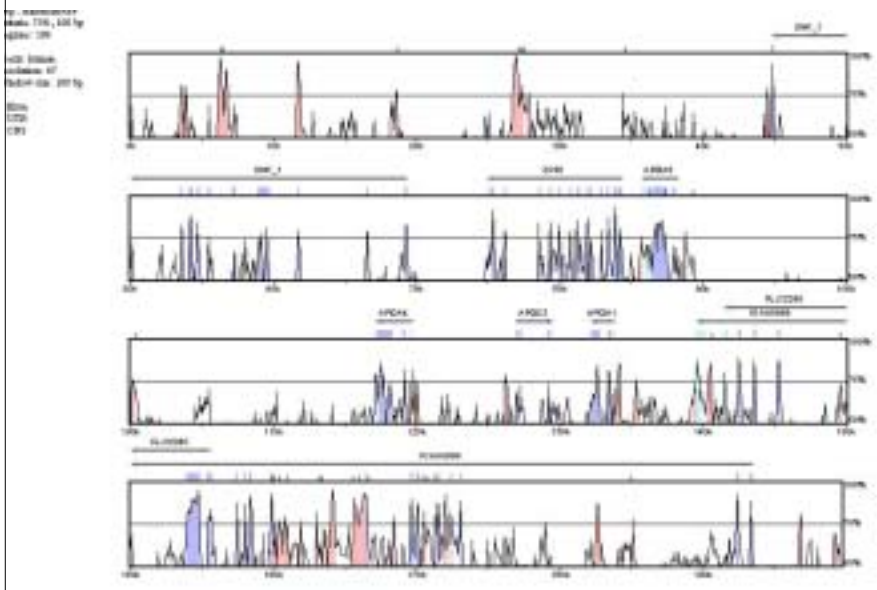
Comparison of Human and Mouse APOA1 Region

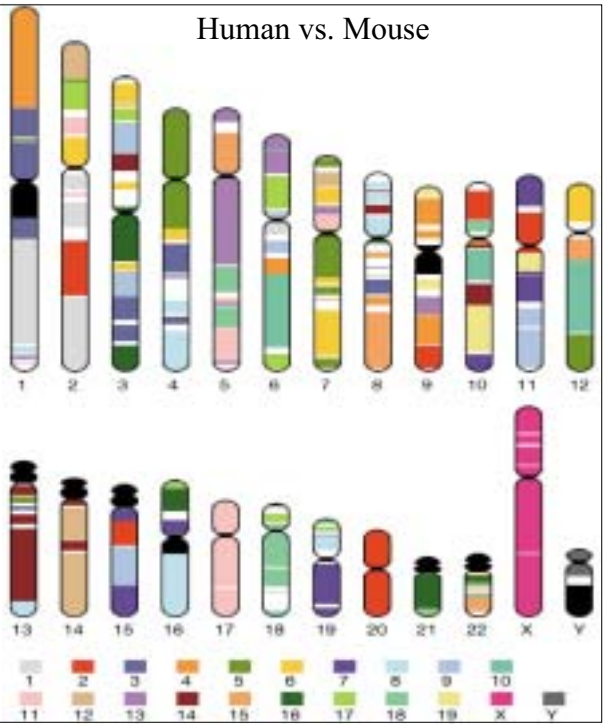


First thing you need:
Synteny:

Preservation of
large (huge)
blocks of chromosome
over evolutionary time,
Preserving gene order

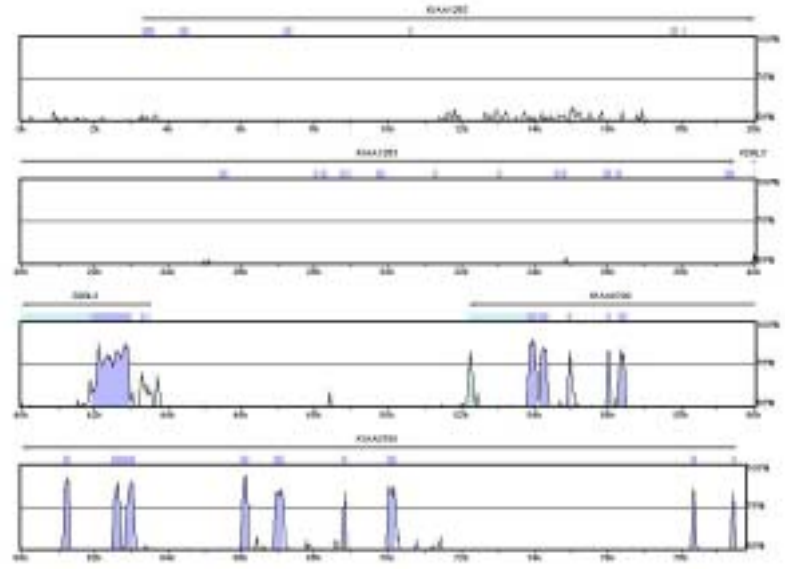
Last count:
Chimp 5 mln years: 1
Mouse 80 mln years: 183
Fugu fish 400-1000s ???

Human vs. Mouse



Known syntenic breakpoint F2RL3

Human
mouse



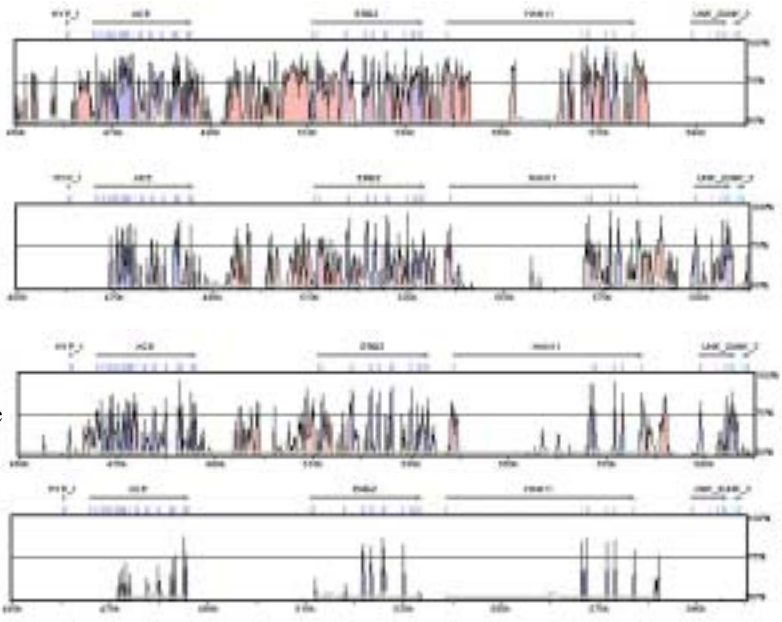
Second thing: proper distance from human

human-lemur

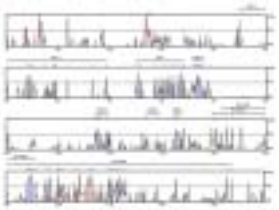
human-rabbit

human-mouse

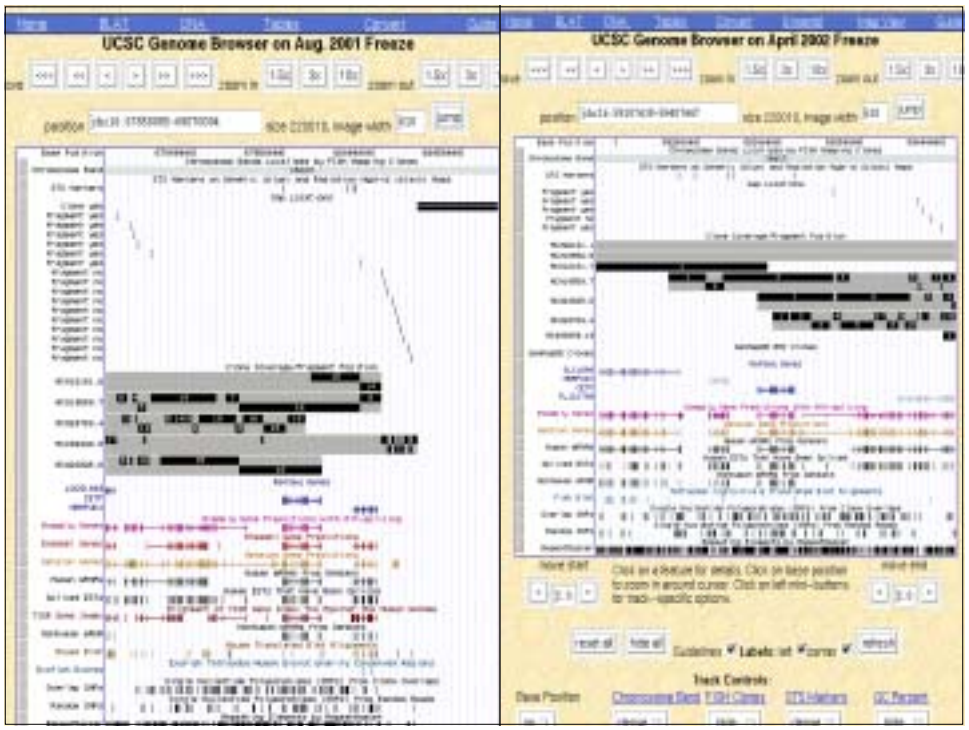
human-chick



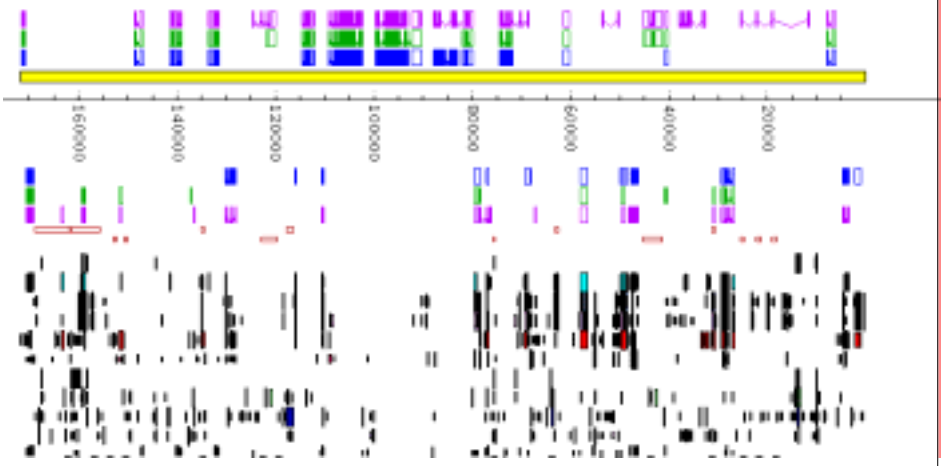
How do we make a vista?



- Need *finished* human sequence
- Human sequence must be annotated
- Need to sequence and assemble second sequence
- Run alignment – get 1 vista
- Total time 1 to 2 weeks
- Godzilla browser- region in seconds
- Is what I do any better?



ACEDB…



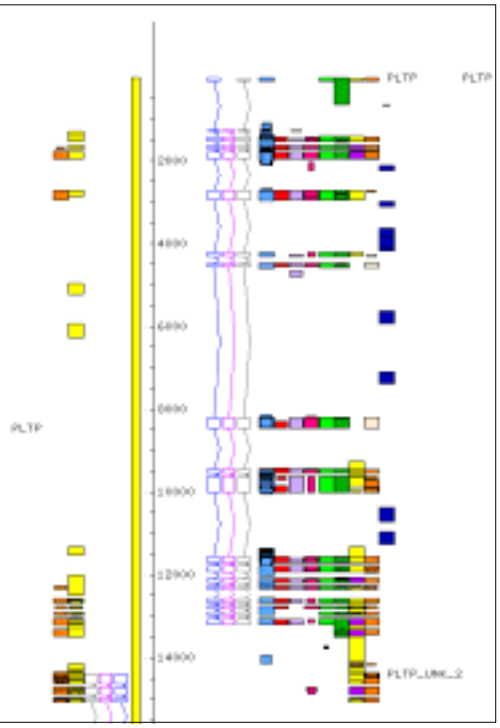
Each box is a different software

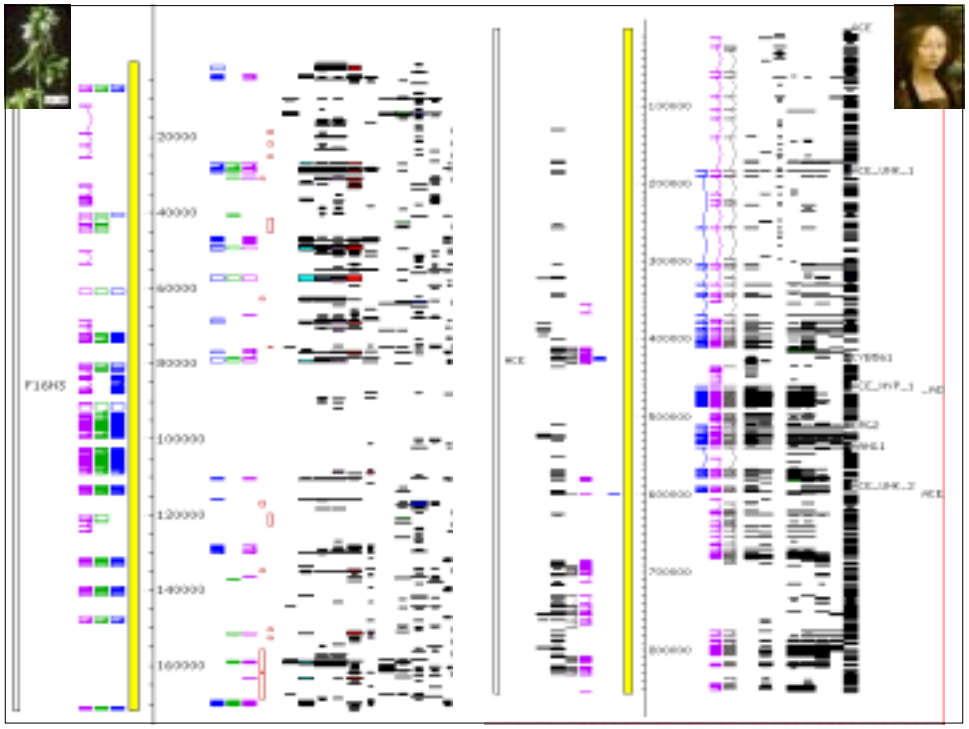
No output is the same

ACeDB requires only ace input

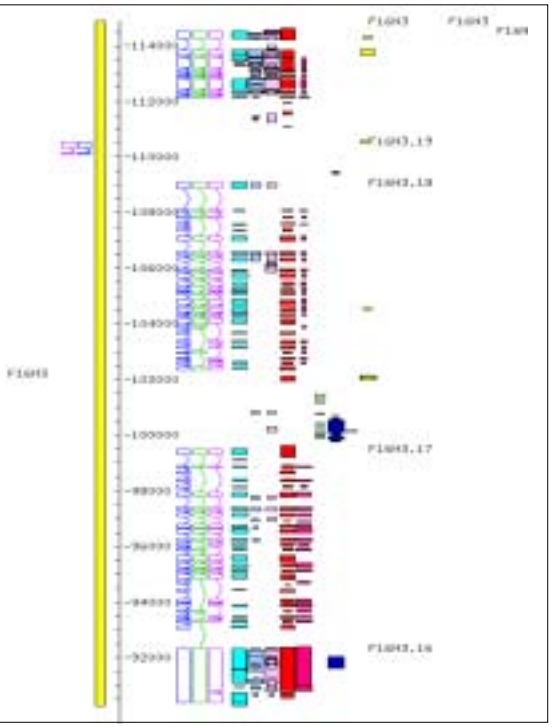
What to do????

PERL!!!!!!!!!!!

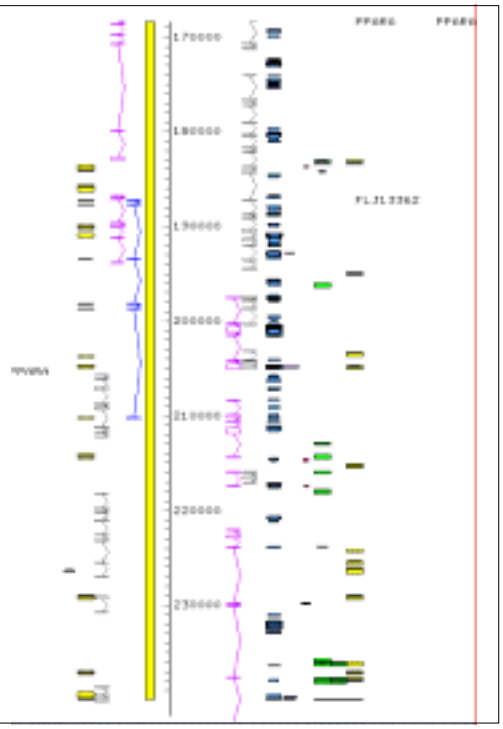




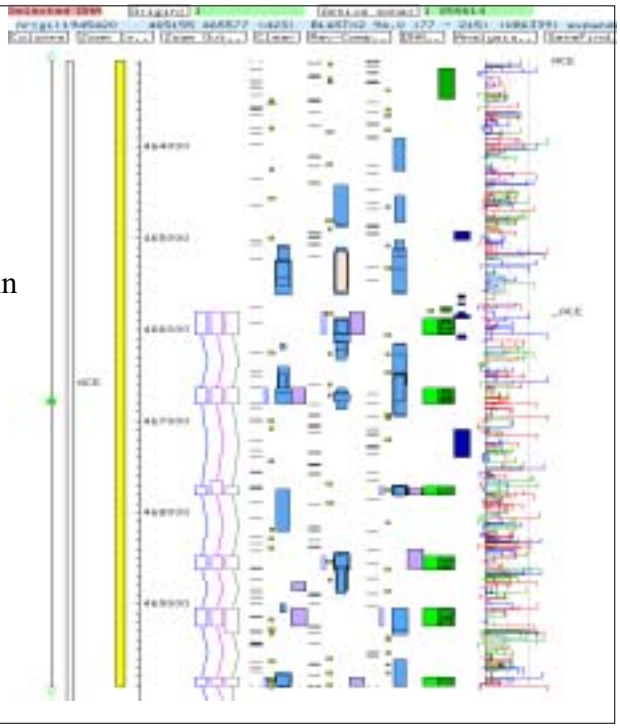
Arabidopsis
No blast evidence
But if it looks like a gene,
It is a gene (Probably).



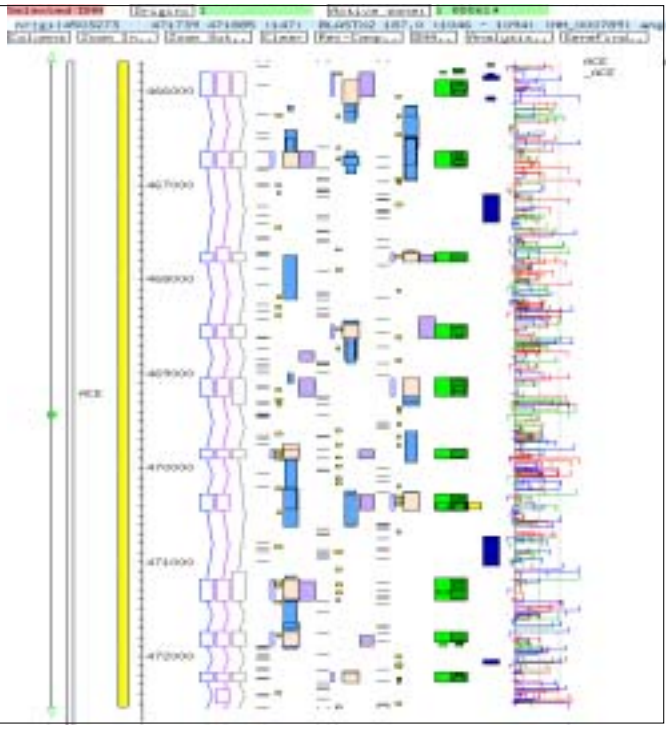
My First Human



Blastx
Sequence translated in
6 frames against the protein
database

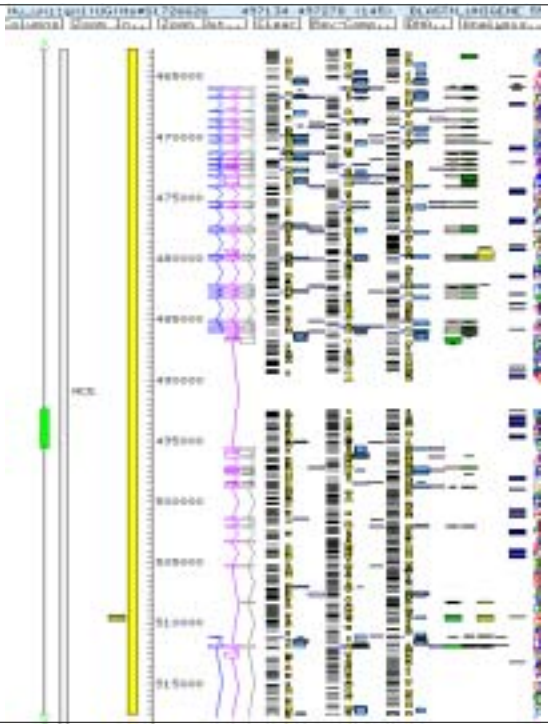


Multiple hits from the same entry are highlighted



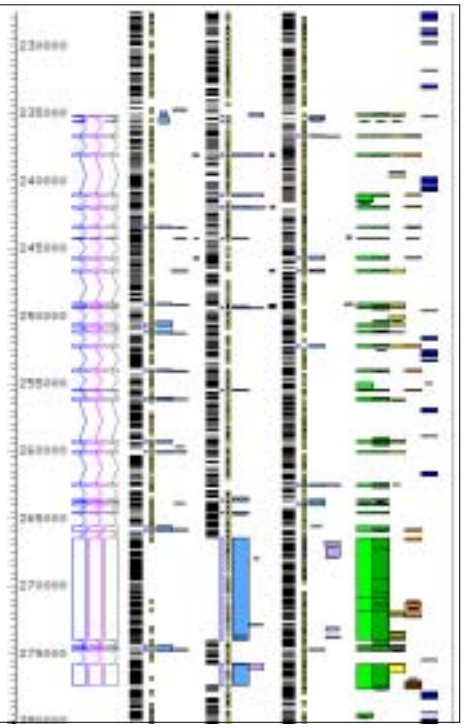
Unigene- superset of Refseq (ncbi curated genes)

Also contains ests that have been clustered into potential genes.



This is not a gene

This is a gene

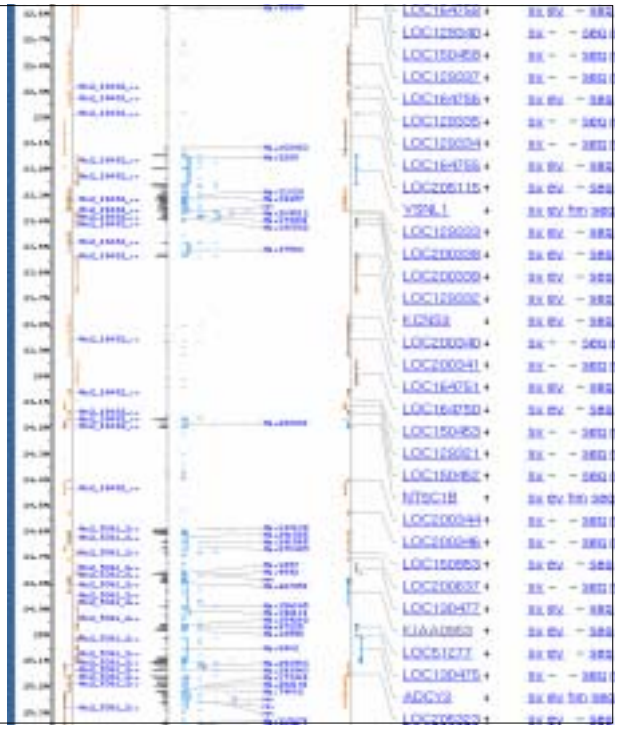


Start Here

Mv = map view

Zoom out from the APOB region…

NCBIs automated annotation-predictions from Genscan



KIAA…/FLJ… cdnas isolated From Japan. Protein product is hypothetical

XM_…

LocusLink:
http://www.ncbi.nlm.nih.gov/LocusLink/

LOC164914

Key for display of mRNAs aligning in this region:

Map/View
Evidence Viewer
Help

- Genomic sequence (C)
- model exons, single (M)
- mRNA exons, single (G, R)
- model exons, overlapping (M)
- mRNA exons, overlapping (G, R)

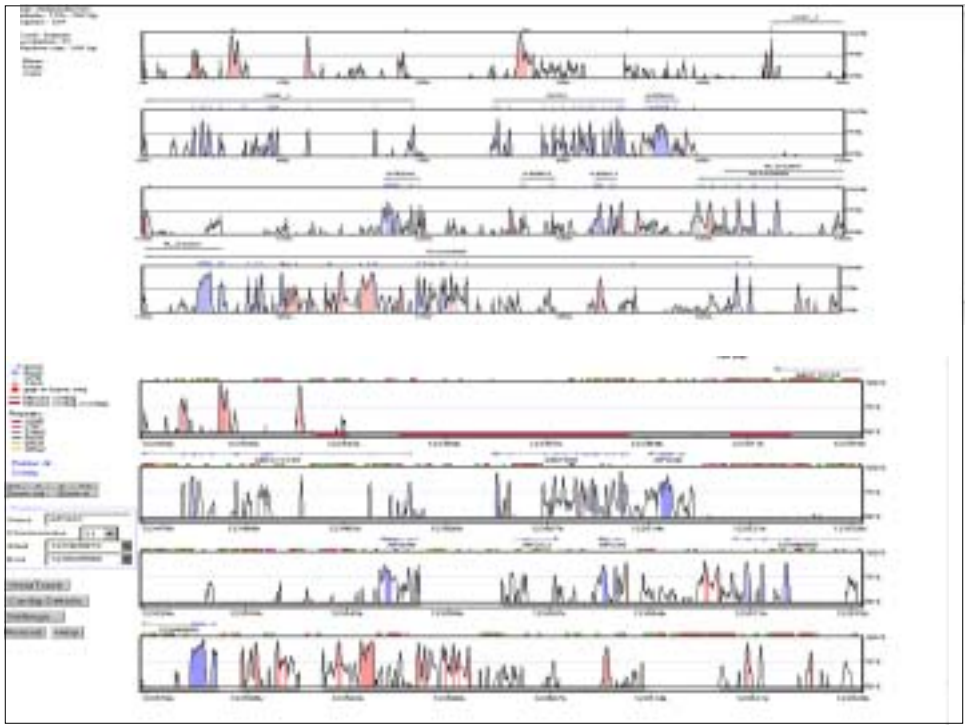
C = config; M = model mRNA; R = RefSeq mRNA; G = GenBank mRNA

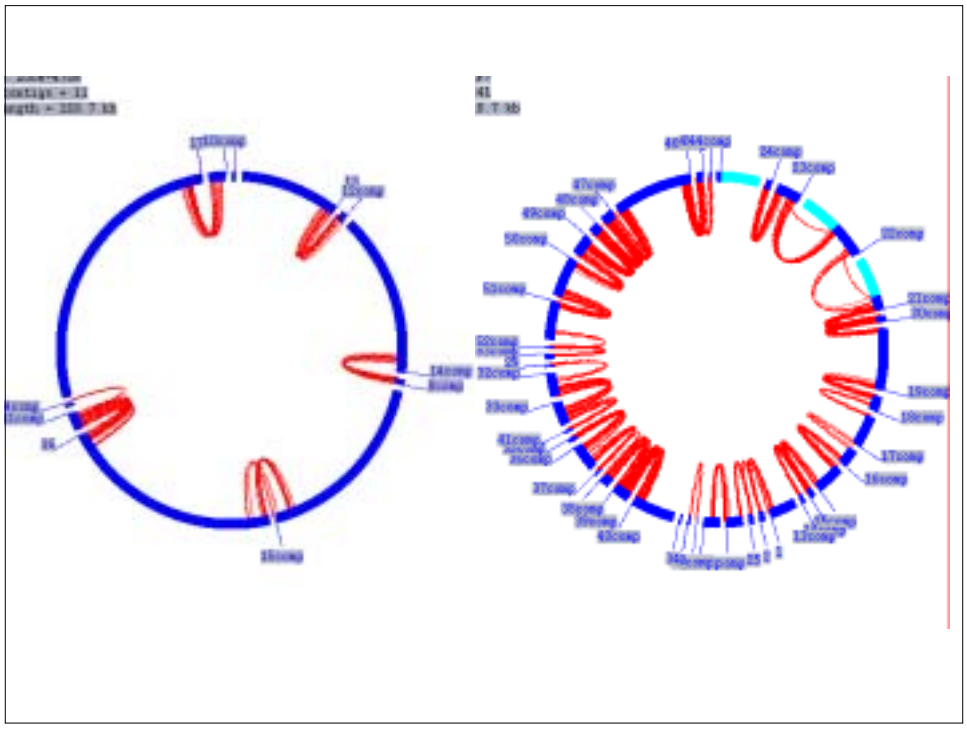
EST density key (E):

- 1 EST
- 2–5 ESTs
- 6–20 ESTs
- 21–88 ESTs
- >100 ESTs

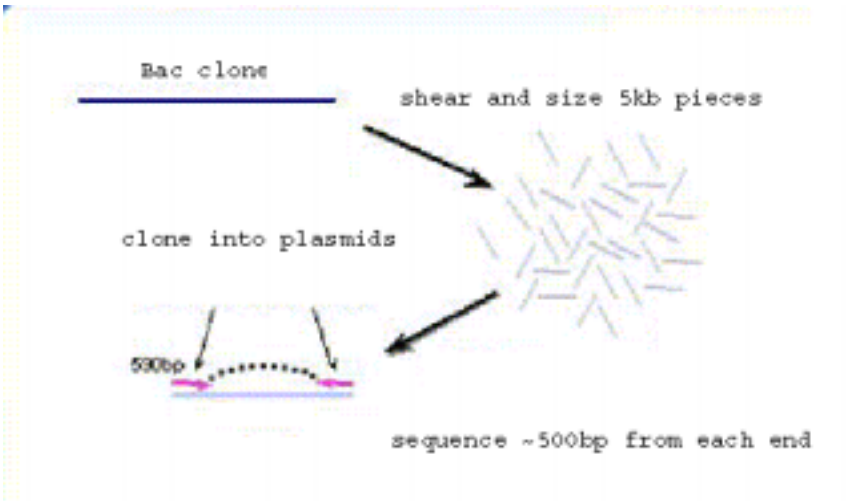
0 exon and 1 gene found in this genomic region spanning 34936 bp.
View graphic only

Mouse over mismatches, indels and unaligned regions to see their exon number.





Sequencing


Bac clone
shear and size 5kb pieces
clone into plasmids
500bp
sequence ~500bp from each end

Assembly: for example Phrap

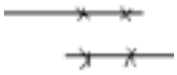
Overlapping reads

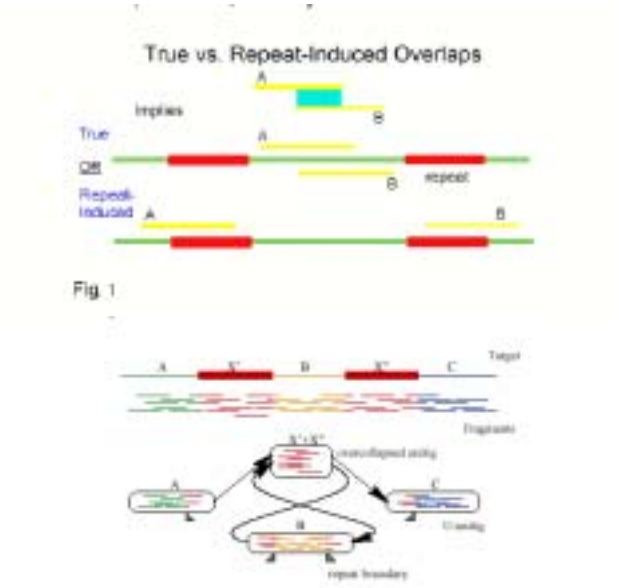
All groups of overlapping sequence make up one contig

Larger structures called supercontigs can be made by connecting the paired-ends over the sequence gaps

These are perfect overlaps.
The computer can figure these out easily.
Simple right?

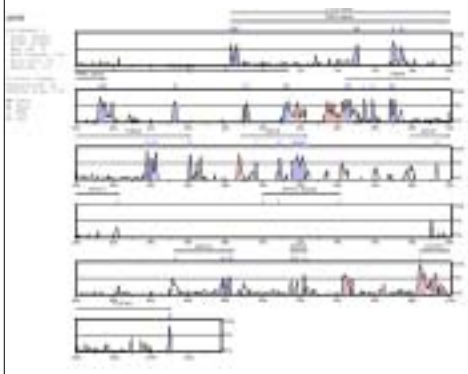
Wrong. Repeats are confused with base call errors.

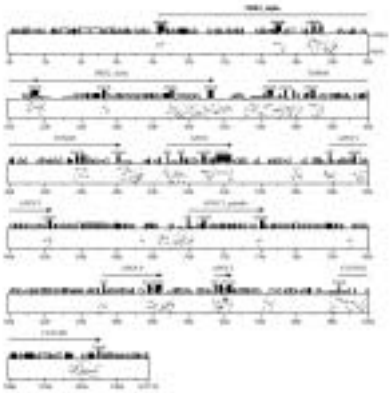




Alignment Method

- GLOBAL- AVID
- LOCAL- Pinmaker





Both use “heuristics” to approximate either a “real” global or local alignment.

To compare 2 sequences, slide them past each other to see where they match maximally, allowing gaps

AGCCUCG
CCGCCAUCG

AGCC UCG
CCGCCAUCG

= 5

Computers can represent this as a matrix:

	A	G	C	C	U	C	G
C			*	*		*	
C			*	*		*	
G		*					*
C			*	*		*	
C			*	*		*	
A	*						
U					*		
C			*	*		*	
G							*

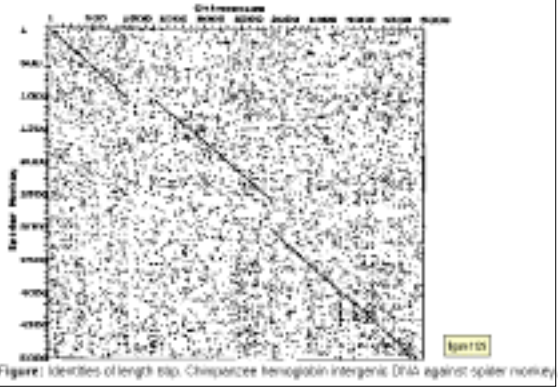


Figure: Identities of length ≥50. Chimpanzee hemoglobin intergenic DNA against spider monkey

Needleman-Wunsch in 5 minutes
The “real thing” global alignment

	M	P	R	C	L	C	Q	R	J	N	C	B	A
P		1											
B												1	
R			1					1					
C				1		1					1		
K													
C				1		1					1		
R								1					
N										1			
J									1				
C				1		1					1		
J									1				
A													1

The math says, take the max
of either:
the diagonal plus the match
the row one down plus the match
the column one down plus the match

Add the 4 plus one for the match = 5

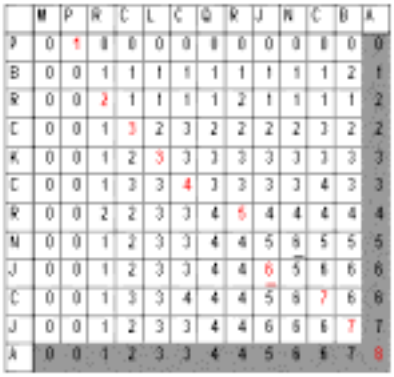
i

	M	P	R	C	L	C	Q	R	J	N	C	B	A
P	0	1	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	1	1	1	1	1	1	1	1	2	1
R	0	0	2	1	1	1	1	2	1	1	1	1	2
C	0	0	1	3	2	2	2	2	2	2	3	2	2
K	0	0	1	2	3	3	3	3	3	3	3	3	3
C	0	0	1	3	3	4	3	3	3	3	4	3	3
R	0	0	2	2	3	3	4	?					
N													
J									1				
C				1		1					1		
J									1				
A													1

j

It looks like: H_{ij} =
$\max(H_{i-1,j-1}) + s(A_i, b_i))$,
$\max(H_{i-k,j-1}) + s(A_i, b_i))$,
$\max(H_{i-1,j-k}) + s(A_i, b_i))$,

You then look for the highest Score. It will always be in the Last row or column. Since you have kept track of where you came from, you can just "traceback" to the other end. This gives you a global alignment of two sequences, it is the maximal score for those two sequences, allowing all gaps.

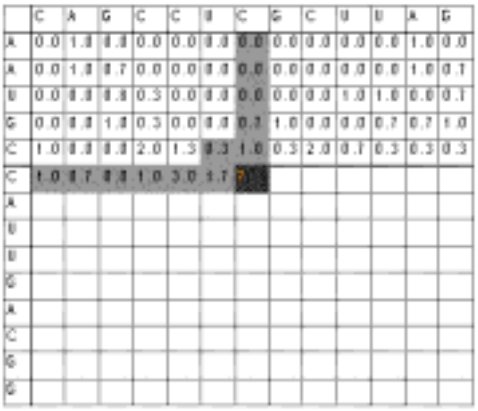


```
MP-RCLCQR-JNCBA
  | ||  |  |  |  |
-PBRCKC-RNJ-CJA
```

Smith –Waterman: Local

Same algorithm, with one exception, you get a penalty when you move off the diagonal

This means that gaps are not free. For each square away, you are Penalized 1/3 in this case.



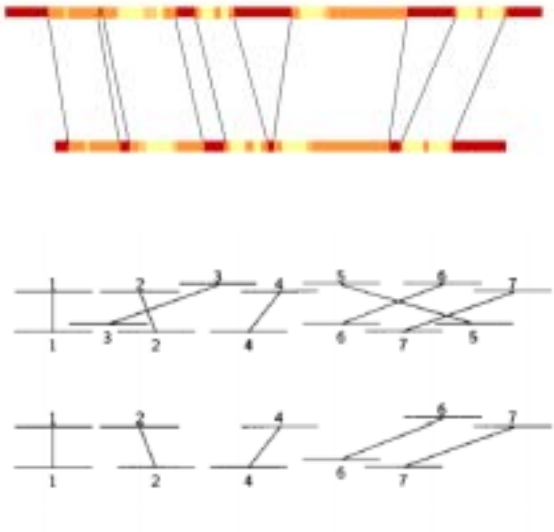
This means that you can get a negative score. If you get a negative score, it stays 0, and you stop.

What happens is that you end up with a maximal score anywhere within the two sequences, you don't have to go to the ends.

	C	A	G	C	C	U	C	G	C	U	U	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.0	2.0	1.7	1.3	2.3	2.7
A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

GCC-UCG
GCCAUUG

Global Heuristic- AVID



Local Heuristic

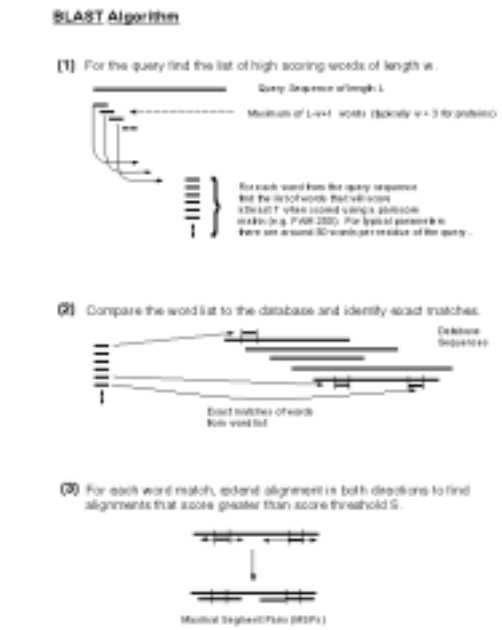
FASTA

⬇

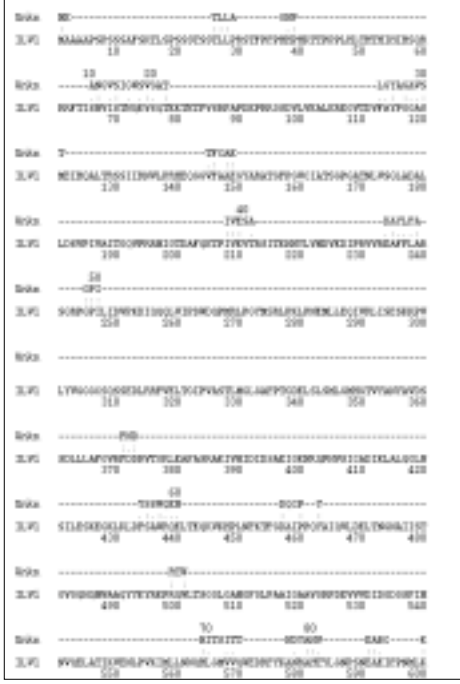
BLAST

⬇

BLAT

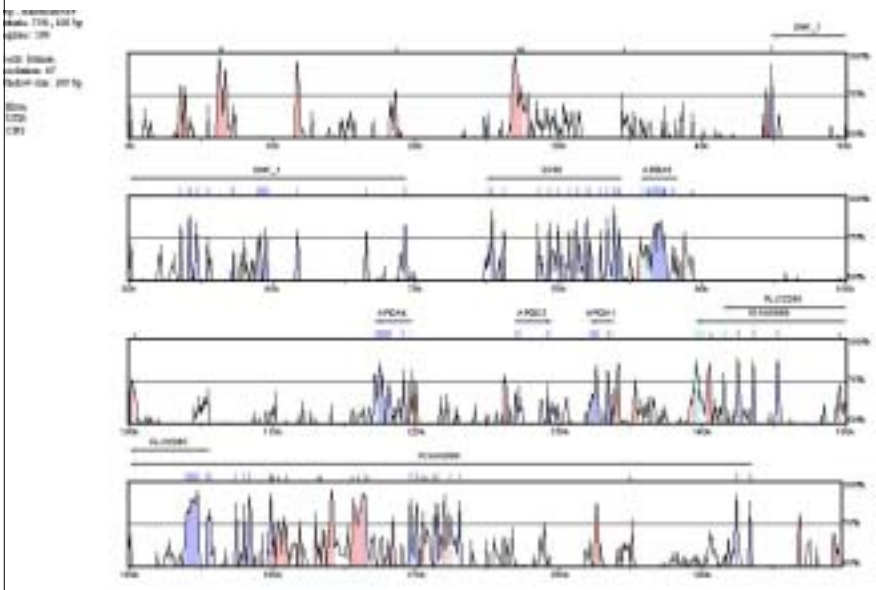


Global



Local

Voila



Aknowledgements:

Eddy Rubin
Inna Dubchak
Jan-Fang Cheng
My friends in the Rubin lab